# CUNY SPS DATA 621 - CTG5 - Final

*Betsy Rosalen, Gabrielle Bartomeo, Jeremy O'Brien, Lidiia Tronina, Rose Koh*

*May 23rd, 2019*

## Contents

# 1 PROJECT DESCRIPTION AND BACKGROUND

## 1.1 Background

[Jeremy's writing this up]

## 1.2 Hypothesis

## 1.3 Our approach, setup, and workflow

[We should discuss this]

# 2 DATA PREPARATION

## 2.1 Cross-validation

[Assuming this requires a little explanation]

## 2.2 Bootstrap surrogata data

[Jeremy's writing this up]

Per the synthpop packace explanation (https://cran.r-project.org/web/packages/synthpop/vignettes/synthpop.pdf): "The basic idea of synthetic data is tovreplace some or all of the observed values by sampling from appropriate probability distributions so that the essential statistical features of the original data are preserved. The approach has been developed along similar lines to recent practical experience with multiple imputation methods although synthesis is not the same as imputation. Imputation replaces data which are missing with modelled values and adjusts the inference for the additional uncertainty due to this process. For synthesis, in the circumstances when some data are missing two approaches are possible, one being to impute missing values prior to synthesis and the other to synthesise the observed patterns of missing data without estimating the missing values. In both cases all data to be synthesised are treated as known and they are used to create the synthetic data which are then used for inference. The data collection agency generates multiple synthetic data sets and inferences are obtained by combining the results of models fitted to each of them. The formulae for the variance of estimates from synthetic data are different from those used for imputed data."

"Our aim in writing the synthpop package (Nowok, Raab, Snoke, and Dibben 2016) for R (R Core Team 2016) is a more modest one of providing test data for users of confidential datasets. Note that currently all values of variables chosen for synthesis are replaced but this will be relaxed in future versions of the package. These test data should resemble the actual data as closely as possible, but would never be used in any final analyses. The users carry out exploratory analyses and test models on the synthetic data, but they, or perhaps staff of the data collection agencies, would use the code developed on the synthetic data to run their final analyses on the original data. This approach recognises the limitations of synthetic data produced by these methods."

## 2.3 Synthesis diagnostics

The original Cleveland dataset contains n = 303 observations over ...

# 3 BUILDING MODELS

Our literature review revealed that of the many approaches that have been taken, certain types of models stand out in terms of their performance.

[Include grid with examples of previous work]

[See notes in literature reviewed and Kaggle projects for suggestions on variable selection and feature engineering for models (links TBC)]

[See also notes on setup and packages for decision tree, random forest, SVM, and Naive Bayes (links TBC]

## 3.1 Logistic regression

## 3.2 Decision tree (CHAID or C&RT?)

## 3.3 Random forest

## 3.4 Support Vector Machines

## 3.5 Naive Bayes

# 4 MODEL REVIEW AND SELECTION

## 4.1 Comparison of performance between models

[sensitivity, specificity, accuracy, others metrics?]

[Between different techniques and between original dataset and synthesize dataset for each technique]

## 4.2 Comparison of performance viz. other studies

# 5 CONCLUSIONS

# 6 APPENDIX

## 6.1 Supplemental tables and figures

## 6.2 R statistical programming code

The appendix is available as script.R file in `projectFinal_heart` folder.

https://github.com/betsyrosalen/DATA_621_Business_Analyt_and_Data_Mining